Modelling habitat distributions for multiple species using phylogenetics (Appendix)

Guillaume Guénard[1]*, Gabriel Lanthier[1]†, Simonne Harvey-Lavoie[1] ‡,
Camille J. Macnaughton[1]§, Caroline Senay[1]¶,
Michel Lapointe[2]‖, Pierre Legendre[1]**and Daniel Boisclair[1]††

23rd August 2016

1. Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre-Ville
Montréal, QC, Canada H3C 3J7

2. Department of Geography
McGill University
Burnside Hall, 805 Sherbrooke W, Room 619
Montreal, QC, Canada H3A 2K6

*guillaume.guenard@gmail.com (corresponding author)
†gabriel.lanthier@umontreal.ca
‡simonne.harvey-lavoie@umontreal.ca
§camille.macnaughton@gmail.com
¶caroline.senay@umontreal.ca
‖Michel.Lapointe@mcgill.ca
**pierre.legendre@umontreal.ca
††daniel.boisclair@umontreal.ca

Field sampling

During the summer months (late June – early September) of the years 2011 – 2013, the Hydronet research group (hereafter simply referred to as "Hydronet") surveyed 28 rivers (15 unregulated and 13 regulated) located in four Canadian provinces (New-Brunswick, Quebec, Ontario, and Alberta). Each river was sampled at multiple locations hereafter referred to as the sampling sites. Sampling sites had 300 m^2 surface area (5 m across by 60 m along the river) and were positioned in alternation (site width was often substantially narrower than river sections), near the left shore (facing downstream), in the middle, and near the right shore of the river, starting randomly, while ensuring that the habitat within them was fairly homogeneous. Locations of the beginning and end of each sampling site were recorded using a GPS unit (GPSMAP® 76sc: Garmin International Inc. 1200 E. 151st Street, Olathe, KS 66062, USA).

Fish sampling consisted in electrofishing (using an LR24® Backpack electrofisher: Smith-Root Inc., 14014 NE. Salmon Creek Ave., Vancouver, WA 98686, USA; electroshocking duration per site: 900 s; mean power: 200 W) and snorkelling surveys performed during daytime hours (between 0830 and 1800). These two sampling methods were used in tandem to minimise the bias associated with because each of these methods shows selectivity toward catching fish of particular species and size classes (Macnaughton et al., 2014). Sampling with both methods involved zigzagging in the upstream direction to cover the whole sampling area. Fish were identified to species or else to the nearest taxonomic level that electrofishers or snorkelers could discriminate visually. Electrofishing was performed by teams of three fishers: one operating the electrofisher and two catching the electroshocked fish with dip nets 38 mm long × 33 mm wide × 20 mm deep, mesh size: 6.35 mm (Smith-Root Inc.). Fish were identified, measured (Total length, TL ±0.1 cm), allowed to recover in cool aerated water, and released at the location where they had been captured. Visual observations were performed by teams composed of two trained snorkelers. Fish with total lengths $\geqslant 3$ cm (fish with total lengths < 3 cm could not be identified reliably by snorkelers) were categorised into size classes, with the first class going from 3 to 5 cm (median size approximately 4 cm), and then in classes 5 cm apart (median sizes: 7.5 cm, 12.5 cm, 17.5 cm, etc.).

Descriptors of the local physical environment that are often regarded as key drivers shaping community structure were estimated in the sampling sites (Knouft et al., 2011; Michel and Knouft, 2014). We estimated water depth (z, in cm, measured using a graduated pole) and velocity (v, in $\mathrm{cm\,s^{-1}}$; Flo-Mate 2000: Marsh-McBirney Inc., 4539 Metropolitan Court Frederick, MD 21704-9452, USA), substrate median grain size (D_{50}, in cm), and the proportion of macrophyte cover (MC, in %) in ten 50 cm × 50 cm plots randomly dispatched within each sampling site (Wolman, 1954; Latulippe et al., 2001). Water temperatures was measured in the field using HOBO UA-002-62 logging thermometers (precision: ±0.5°C; Onset Computer Corp., 470 MacArthur Blvd., Bourne, MA 02532, USA; 3-10 devices / river). Devices were deployed in riffle, run, or shallow pool habitats and set to record water temperature at 15 min intervals for at least 9 weeks. We used the number of heating degree-days (DD, in °C d) as an environmental descriptor, which we estimated as the sum of mean daily water temperatures above 0°C for a period of nine weeks encompassing the four weeks preceding the hottest week of the summer, and the four weeks following it (between late-May to late-September). During the low summer flows and on days without rain, we took between one and nine water samples in the main stream of the rivers. These samples were collected in 250 mL acid-washed, high-density polyethylene bottles (Nalgene®, Nalge Nunc International Corporation), kept at 4 °C, and shipped to the University of Alberta's Biogeochemical Laboratory for analysis. Total phosphorus (TP, in $\mu\mathrm{g\,L}^{-1}$) was determined using persulfate oxidation (Ng, 2010). For each river, individual TP values that differed by 15 standard deviations or less from the river mean were retained; we took the average of the latter as the total phosphorus for the whole river.

Data processing

In the present study, we used fish density ($\mathrm{fish\,(100\,m^2)^{-1}}$) by species and size class as the response variable of the phylogenetic habitat model. To obtain a *species* × *location* response matrix that was not overly sparse (i.e. did not contain too many zeros) as well as to facilitate computations, modelling was performed on the basis of whole rivers. For each river, values of z, v, D_{50}, and MC were averaged. Mean densities for whole rivers were estimated separately for each combination of species and size class using Generalised Linear Models (GLM: Hastie and Pregibon, 1991; Poisson distributed) fitted with the values of depth, velocity and substrate grain size for each sampling site as descriptors. We took the fitted values of each GLM (one per river for each combination of species and size-class), estimated for the average environmental conditions observed in the river, as the mean fish density. Species and size class combinations that were not observed in a given river were assumed to be absent from that river and assigned a density value of $0\,\mathrm{fish\,(100\,m^2)^{-1}}$.

For any given species, we discarded all size classes for which no fish was observed, and merged the size classes for which fish were present in two rivers or less. Size class merging was performed by coalescing size classes with insufficient numbers of observations with an adjacent size class. When two adjacent size classes were present on either side of the class with low number of observations, the one with the smallest number of observations was chosen. The mean size of the newly formed size class was taken as the mean of their median sizes weighted by their respective numbers of observations. Size class merging stopped when all those remaining had been observed in two rivers or more. The species represented by a single size class that were present in only one river were discarded.

We used the phylogeny published by Hubert et al. (2008), which was estimated using K2P distances (Kimura, 1980) calculated from a standard 652 base pairs «bar code» region of the Cytochrome Oxydase subunit I (COI) mitochondrial gene. That tree was used to calculate Phylogenetic eigenvector maps (PEM; Guénard et al., 2013). PEM required a tree tip for every combination of species and size class. Therefore, every single tree tip representing species with more than one size class was merged into a star tree having branch lengths of 0 and as many tips as the species had size classes. Since the merged star tree had branch lengths of 0, the number of phylogenetic eigenvectors with non-zero eigenvalues obtained from the approach remain one minus the number of species, with each eigenvector defined for all the species and size classes; they are, however, invariant across observations involving the same species.

We used spatial eigenvector maps (SEM; Borcard and Legendre, 2002; Dray et al., 2006; Diniz-Filho et al., 2013; Legendre and Legendre 2012, Chapter 14) to model patterns of spatial variation among the rivers and account for spatial auto-correlation (Legendre, 1993; Dormann et al., 2007). These eigenfunctions were calculated from the geodesic distances among rivers.

In the present study, the design matrix $\mathbf{X}$ contained the species descriptors, which included the median standard length of the fish in each size class as a species trait in addition to the PEM descriptors that represented the among-species phylogenetic patterns of trait variation. It is noteworthy that, in the model, variation among species and size classes was explained in part by the median fish size and the PEM, because fish of different species had different median sizes, on average, but within-species variation was only associated to variation in the median fish size, because PEM were constant across size classes for a given species. The design matrix $\mathbf{Z}$ of site descriptors included environmental variables and the spatial eigenfunctions. Since two types of row descriptors were used with two types of column descriptors, there were eight types of bilinear model terms. Among them, four were modelling main effects acting alone in the model: 1) the traits, 2) the phylogeny (PEM), 3) the environment, 4) space (spatial eigenvectors). There were also four second-order interaction terms: 1) trait-environment, 2) trait-space, 3) phylogeny-environment, 4) phylogeny-space. Therefore the first two columns of matrix $\mathbf{R}$ represent the different types of row descriptors, namely *trait* (fish total length; column 1) and *phylogeny* (PEM; column 2), the next two columns represent the column descriptors, namely *environment* (z, v, D_{50}, MC, DD, and TP; column 3) and *space* (SEM; column 4), and the last four columns of $\mathbf{R}$ represent their interactions (column 5: *trait* $\wedge$ *environment*, column 6: *trait* $\wedge$ *space*, column 7: *phylogeny* $\wedge$ *environment*, and column 8: *phylogeny* $\wedge$ *space*). We fitted the model using a Poisson-distributed Generalized Linear Model (GLM).

Details on model estimation

To ensure that the resulting model was general, rendered dependable predictions, and avoided over-fitting, we regularised the bilinear regression model using elastic net regularisation (Zou and Hastie, 2005), which is a combination of the Least Absolute Shrinkage and Selection Operator (LASSO; Tibshirani, 1996; the L_1 regularisation norm) and Tikhonov regularisation (Tikhonov and Arsenin, 1977, the latter being also known as ridge regression; the L_2 regularisation norm). That method involves estimating matrix $\mathbf{B}$ as the solution minimising an objective function f involving size of the (standardised) regression coefficients as follows:

$$f_{\alpha,\lambda,\xi}(y;x,z|\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m}\varepsilon_{i,j}^2 + \lambda\sum_{k=1}^{p}\sum_{l=1}^{q}\xi_{k,l}\left(\alpha|\beta_{k,l}| + \frac{(1-\alpha)}{2}|\beta_{k,l}|^2\right) \tag{1}$$

where α is a parameter varying between 0 and 1, which controls the relative importance of the regularisation norms, λ is the average penalty of all the descriptors ($0 \leq \lambda < \infty$), and $\xi_{k,l}$ are penalty factors specific to particular terms of the model, $\beta_{k,l}$ are the standardised regression coefficients, and $\varepsilon_{i,j}$ are the residuals of the regression model. Therefore, any variable with $\lambda\xi_{k,l} > 0$ (the minimum value) experiences an effective shrinkage of its contribution to the model. In that framework, parameter λ controls the average amount of shrinkage for all the regression coefficients whereas parameters $\xi_{k,l}$ allows one to attribute greater shrinkage to certain groups of regression coefficients with respect to others. That feature can be exploited for the purpose of maximising the predictive power. We estimated α, λ, and $\xi_{k,l}$ using embedded models estimated by cross-validation as those maximising the predictive power of the models. Parameters α and λ were estimated as constants $\mathrm{c}_\alpha = \mathrm{logit}\,\alpha$ and $c_\lambda = \log\,\lambda$, respectively, to ease optimisation (since the domain of c_α and c_λ need not be bounded), whereas parameters $\xi_{k,l}$ were estimated through a logistic model with a factorial design that was defined as follows:

$$\mathbf{R}[c_\xi] = [\mathrm{logit}\,\xi_{k,l}], \tag{2}$$

where $\mathbf{R}$ is a binary design matrix with $(p \cdot q - 1)$ rows, each of them representing the columns of $\mathbf{Z} \otimes \mathbf{X}$, with the exception of the model intercept (the first column), which is assigned a constant penalty factor of 0.5, and $(r+1)\,(s+1) - 1$ columns, where r is the number of types of row descriptors and s is the number of types of column descriptors, while $[c_\xi]$ is a column vector of the regularisation model coefficients. It is noteworthy that following Eq.2, penalty values are equal for all the variables involved in the same model term. We chose a logistic rather than a linear model to represent the penalty factors because constraining the penalty factors between 0 and 1 avoids interference with the estimation of c_λ during the optimisation process. The columns of matrix $\mathbf{R}$ represent the different types of row descriptors, column descriptors, as well as their interactions. Elements in each column of $\mathbf{R}$ take the value 1 whenever the descriptor is involved in the type of descriptor it represents, or their interactions, and 0 otherwise.

For phylogenetic habitat modelling, there will be at least one type of each of the row and column descriptors (i.e. $r = 1$ and $s = 1$), namely a set of variables representing the phylogeny, and a set of variables representing the environment. In the simplest scenarios, matrix $\mathbf{R}$ will therefore have three columns: one that represents the marginal effect of the phylogeny, one that represents the marginal effect of the environment, and a third representing the interaction between phylogeny and the environment. In such a scenario, rows of $\mathbf{R}$ corresponding to the columns of $\mathbf{Z} \otimes \mathbf{X}$ representing the marginal effect of the the phylogeny will take the value 1 in the first column of $\mathbf{R}$ and 0 elsewhere. On the other hand, rows of $\mathbf{R}$ corresponding to the columns of $\mathbf{Z} \otimes \mathbf{X}$ representing the phylogeny-environment interaction will also take the value 1 in the first column of $\mathbf{R}$, but also in the second, which represents the environment, and in the third column, which represents the phylogeny-environment interaction, thereby allowing the penalties applied to the marginal effects of phylogeny and environment to be non-additive with that applied to their interactions. In practice, it may be useful to use traits alongside the variables describing the phylogeny as row descriptors and variables describing spatial variation alongside environmental variables as column descriptors.

Assessing the effect of flow regulation

We assessed the effect of flow regulation on fish density by subtracting the predicted fish density values by species and size classes, obtained using the unregulated river model, from the fish densities observed at the sampling sites. The density differences thus obtained were negative when the model predicted higher densities than that observed in the field; positive values were obtained for observed densities that were higher than those predicted. Prior to further analysis, we $\log(x + 1)$-transformed these density differences on the basis of their absolute values while conserving their signs as follows:

$$x_{transfromed} = \mathrm{sign}(x_{original})\ \log(|x_{original}| + 1), \tag{3}$$

where $x_{original}$ and $x_{transformed}$ are the original and transformed differences in fish density, respectively, and function $\mathrm{sign}(x)$ returns -1 when $x < 0$, 0 when $x = 0$, and 1 when $x > 0$. The necessity for that transformation was consequential to the fact that a Poisson distribution was assumed for modelling. It was used to mitigate over-dispersion and avoid conclusions driven only by a few extreme values.

Data collected

A total of 989 sites were sampled in the 28 rivers but because of malfunctioning equipment during field sampling, information about water velocity was not available for 48 sites located on four rivers (St-Jean: 17 sites, Elbow: 12, Kananaskis: 10, and Petit-Saguenay: 9); these sites were thus discarded, leaving us with 941 usable sites for modelling. Mean river environmental conditions and species densities were calculated using numbers of sites ranging from 16 (in Petit-Saguenay) to 50 (in Au saumon and Bécancour). Water depth was left-skewed and was log-transformed before further treatment as were water velocity and substrate grain size; both were $\log(x + 1)$-transformed because values of 0 were present. Water depth (z) ranged 25.8–55.1 cm (mean= 36.82 cm), v ranged 0.07–0.59 $\mathrm{m\,s^{-1}}$ (mean= 0.296 $\mathrm{m\,s^{-1}}$), D_{50} ranged from 0.29 to 38.68 cm (7.057 cm), MC ranged from 0 to 21%, DD ranged from 367 to 1540 °C d, and TP ranged from 1.00 to 3.42 $\mu\mathrm{g\,L^{-1}}$ among the rivers. A total of 244 combinations of species and size classes, involving 61 species, were observed in at least one of the 28 rivers. After merging the sparsely observed size classes, 143 combinations of species and size classes, involving 48 species, were retained (see Fig. 1 in the main text). On average, each species was represented by 2.6 size classes, with numbers ranging from one (six species) to nine (one species: the white sucker, *Catostomus commersoni*). The characteristic (median) sizes of these classes ranged from 4.0 to 42.5 cm (mean: 9.50 cm). The sample size ($n\,p$) for the present application scenario was 2145 observations (143 combinations of species and size class sampled in the 15 unregulated rivers).

Details on the interactions between phylogeny and the environment

We found a total of 22 interactions terms between phylogenetic eigenfunctions and five of the six environmental descriptors (v, D_{50}, DD, TP, and MC). Mean current velocity interacted with one phylogenetic eigenfunction (PE36) and described the Common shiner (*Luxilus cornutus*), the Bluntnose (*Pimephales notatus*), and the Fathead (*P. promelas*) minnows as being more abundant in rivers with faster current than are the eastern shiners (i.e. species of genus *Notropis*). Substrate grain size interacted with five phylogenetic eigenfunctions (PE12, PE17, PE27, PE30, and PE36); their overall effect indicated that, in the family Cyprinidae, riffle daces (i.e. the Longnose dace, *Rhinichthys cataractae*, and the Blacknose dace, *R. atralutus*) tended to be more abundant in rivers with coarser substrate than the average, while the Fallfish (*Semotilus corporalis*), the Bluntnose and fathead (genus *Pimephales*) minnows, the eastern shiners, and the Golden shiner (*Notemigonus crysoleucas*) tended to be abundant in rivers with finer substrate. Temperature interacted with three phylogenetic eigenfunctions (PE17, PE19, and PE35), their overall contributions indicating that the White sucker (*Catostomus commersoni*), the redhorses (i.e. the Silver redhorse, *Moxostoma anisurum*, and the Shorthead redhorse, *M. macrolepidotum*), the Lake chub (*Couesius plumbeus*), and the Creek chub (*Semotilus atromaculatus*) were observed in greater densities in cold environments whereas the Longnose dace, the Fallfish, the White sucker, the Brown bullhead (*Ameiurus nebulosus*), and the Stonecat (*Noturus flavus*) were denser in warmer environments. Total phosphorus interacted with a total of 11 phylogenetic eigenfunctions and their combined effects were associated to a wide range of differential species responses to TP, which we enumerate as follows:

1. The complex formed of the Johnny darter and the Tessellated darter (*Etheostoma olmstedi nigrum - E. olmstedi*; these species could not be discriminated by phylogenetic analysis) is found in higher densities in high TP environments, while the Fantail darter (*E. flabellare*) is denser in lower TP environments. Two phylogenetic eigenfunctions were involved in that interaction: PE33 and PE42.

2. Among the Percidae, the Walleye (*Sander vitreus*) appears to be favoured in high TP conditions whereas the Yellow perch (*Perca flavescens*) and two roughbelly darters (Genus *Percina*: the Common logperch, *P. crapodes*, and the Channel darter, *P. copelandi*) were denser in lower TP environments. PE23 and PE45 were involved in that interaction.

3. Among the eastern shiners, the Sand shiner (*Notropis stramineus*) is more prevalent in high TP in comparison with the Mimic shiner (*N. volucellus*). PE44 was involved here.

4. In genus *Semotilus*, higher TP is concomitant with higher densities of the Fallfish compared to the Creek chub and the Lake chub. PE17 and PE22 were behind that interaction.

5. The riffle daces and the Cutlips minnow (*Exoglossum maxillingua*; the effect was described by PE29) as well as the Fathead minnow (the effect was, this time, associated with PE30 and PE40) seem to find high TP environments more favourable than the other Cyprinids.

6. Among the Catostomidae, the White sucker are found to be denser in high TP environments while the Longnose (*Catostomus catostomus*) and Mountain (*C. platyrhynchus*) suckers as well as the redhorses appear to be denser in rivers with lower TP. Here, PE19 and PE45 were involved.

Macrophyte cover interacted with three phylogenetic eigenfunctions; their combined effects described the riffle daces, the Cutlips minnow, the Golden shiner, and the Fathead minnow as being found in higher densities in rivers with low MC, compared to the eastern shiners, the Common shiner (*Luxilus cornutus*), the Fallfish and the Bluntnose minnow, which had higher densities in rivers with higher MC. These examples illustrate the capacity of the method described in this paper to represent, in a single model, many details about the habitat requirements of many species, using knowledge of their common evolutionary history, and how these factors drive community structure.

Analysis data and script

The raw data and analysis script used for the exemplary scenario are available online at `http://dx.doi.org/10.5061/dryad.60n52`.

Tables

Table 1: Bilinear model coefficients estimated by the elastic net procedure.

	$Int.$	TL	PE_1	PE_{12}	PE_{17}	PE_{19}	PE_{22}	PE_{23}	PE_{27}
$Int.$	-3.58	—	-8.79	—	—	—	—	—	—
$\log z$	—	-0.0193	—	—	—	—	—	—	—
$\log(v+1)$	—	-0.0837	—	—	—	—	—	—	—
$\log(D_{50}+1)$	—	-0.0158	—	-1.11	0.107	—	—	—	0.72
DD	0.00131	-4.4e-05	—	—	0.000964	0.000785	—	—	—
$\log TP$	1.79	—	—	—	0.725	0.95	0.441	0.987	—
MC	—	-0.0242	—	0.395	—	—	—	—	—
SE_1	0.313	-0.0316	—	—	—	—	—	—	—
SE_2	—	0.103	—	—	—	—	—	—	—
SE_3	—	-0.0777	—	—	—	—	—	—	—
SE_4	—	0.0238	—	—	—	—	—	—	—
SE_7	—	-0.00938	—	—	—	—	—	—	—
SE_8	—	0.102	—	—	—	—	—	—	—
SE_{11}	—	-0.00975	—	—	—	—	—	—	—
SE_{12}	—	0.0272	—	—	—	—	—	—	—
SE_{13}	—	0.0679	—	—	—	—	—	—	—
SE_{14}	—	-0.0182	—	—	—	—	—	—	—

	PE_{29}	PE_{30}	PE_{33}	PE_{35}	PE_{36}	PE_{40}	PE_{42}	PE_{44}	PE_{45}
$Int.$	—	—	—	—	—	—	—	—	—
$\log z$	—	—	—	—	—	—	—	—	—
$\log(v+1)$	—	—	—	—	4.8	—	—	—	—
$\log(D_{50}+1)$	—	0.127	—	—	0.318	—	—	—	—
DD	—	—	—	-0.000554	—	—	—	—	—
$\log TP$	-0.224	0.0298	-2.5	—	—	-0.295	0.256	0.624	-0.165
MC	—	—	—	—	—	-0.105	—	—	—
SE_1	—	—	—	1.91	—	—	—	—	—
SE_2	—	—	—	—	—	—	—	—	—
SE_3	—	—	—	—	—	—	—	—	—
SE_4	—	—	—	—	—	—	—	—	—
SE_7	—	—	—	—	—	—	—	—	—
SE_8	—	—	—	—	—	—	—	—	—
SE_{11}	—	—	—	—	—	—	—	—	—
SE_{12}	—	—	—	—	—	—	—	—	—
SE_{13}	—	—	—	—	—	—	—	—	—
SE_{14}	—	—	—	—	—	—	—	—	—

$Int.$: intercept, TL: total length (cm), PE_x: x^{th} phylogenetic eigenvector, z: depth (cm), v: velocity ($\text{cm}\,\text{s}^{-1}$), D_{50}: median grain size (cm), DD: cumulative number of degree-day (°C d), TP: total phosphorus ($\mu\text{g}\,\text{L}^{-1}$), MC: percent macrophyte cover (%), SE_x: x^{th} spatial eigenvector, —: placeholder for coefficients estimated to numerically 0; all logarithms are base e

References

Borcard, D. and Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Model.*, 153:51–68.

Diniz-Filho, J. A. F., Diniz, J. V. B. P. L., Rangel, T. F., Soares, T. F., de Campos Telles, M. P., Garcia Collevatti, R., and Bini, L. M. (2013). A new eigenfunction spatial analysis describing population genetic structure. *Genetica*, 141:479–489.

Dormann, C. F., McPherson, J. M., Araújo, M, B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., and Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30:609–628.

Dray, S., Legendre, P., and Peres-Neto, P. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbor matrices (pcnm). *Ecol. Modelling*, 196:483–493.

Guénard, G., Legendre, P., and Peres-Neto, P. (2013). Phylogenetic eigenvector maps (PEM): a framework to model and predict species traits. *Meth. Ecol. Evol.*, 4:1120–1131.

Hastie, T. J. and Pregibon, D. (1991). *Generalized linear models*, volume Statistical models in S, chapter 6, pages 195–247. Wadsworth, Pacific Grove, CA.

Hubert, N., Hanner, R., Holm, E., Mandrak, N. E., Taylor, E., Burridge, M., Watkinson, D., Dumont, P., Curry, A., Bentzen, P., Zhang, J., April, J., and Bernatchez, L. (2008). Identifying canadian freshwater fishes through DNA barcodes. *PLOS ONE*, 3:e2490.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120.

Knouft, J. H., Caruso, N. M., Dupre, P. J., Anderson, K. R., Trumbo, D. R., and Puccinelli, J. (2011). Using fine-scale gis data to assess the relationship between intra-annual environmental niche variability and population density in a local stream fish assemblage. *Meth. Ecol. Evol.*, 2:303–311.

Latulippe, C., Lapointe, M. F., and Talbot, T. (2001). Visual characterization technique for gravel-cobble river bed surface sediments; validation and environmental applications contribution to the programme of CIRSA (Centre Interuniversitaire de Recherche sur le Saumon Atlantique). *Earth Surf. Process. Landforms*, 26:307–318.

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology*, 74:1659–1673.

Legendre, P. and Legendre, L. (2012). *Numerical Ecology, 3rd English edition.* Elsevier Science B.V., Amsterdam, The Netherlands.

Macnaughton, C. J., Harvey-Lavoie, S., Senay, C., Lanthier, G., Bourque, G., Legendre, P., and Boisclair, D. (2014). A comparison of electrofishing and visual surveying methods for estimating fish community structure in temperate rivers. *River Res. Appl.*, 31:1040–1051.

Michel, M. J. and Knouft, J. H. (2014). The effects of environmental change on the spatial and environmental determinants of community-level traits. *Landscape Ecol.*, 29:467–477.

Ng, J. (2010). Subsampling water samples in the field. Technical report, Biogeochemical Analytical Laboratory,University of Alberta, Edmonton, AB, Canada.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, 58:267–288.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solution of Ill-posed Problems.* Winston & Sons, Washington, DC, USA. ISBN 0-470-99124-0.

Wolman, M. G. (1954). A method of sampling coarse river-bed material. *Trans. Am. Geophy. Union.*, 35:951–956.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, 67:301–320.